图书情报工作
第 65 卷 第 14 期　2021 年 7 月

基于题录信息的领域学术文献细粒度分类方法研究*

雷兵[1,2]　刘小[1,2]　钟镇[1,2]

[1] 河南工业大学管理学院　郑州 450001　[2] 河南工业大学商务智能与知识工程实验室　郑州 450001

摘　要：[目的/意义]针对领域学术文献，基于题录信息构建按照“研究内容”与“研究方法”的双标签分类模型，为学术文献的细粒度分类提供方法借鉴。[方法/过程]以深度学习中卷积神经网络为基础模型，将题名、摘要、关键词、刊名、作者、机构等题录信息分为显性特征和隐性特征，通过显性特征提取、隐性特征映射等步骤，形成特征词数组，在此基础上生成词向量矩阵，经过卷积层、池化层与 Softmax 层处理，完成分类任务。[结果/结论]以电子商务领域文献为例进行实验验证，结果显示，该模型按“研究内容”与“研究方法”双标签分类的宏 F_1 值分别为 0.74、0.81，不仅明显优于传统机器学习方法，也比仅使用显性特征的深度学习分类方法高。

关键词：学术文献　主题分类　题录信息　深度学习　卷积神经网络

分类号：G250

DOI：10.13266/j.issn.0252－3116.2021.14.015

1　引言

学术文献的主题分类是图书情报学领域的一项重要的基础工作[1]，不仅能提高学者学术信息检索效率，还可以帮助科技管理、文献管理平台等机构更为准确地解析领域发展方向，制定更为合理的政策或规则，进而推进科技工作快速发展[2]。但近年来，随着科研领域的不断细化以及学术文献数量快速增长，基于传统的文献题录信息和手工或简单机器学习的分类方法暴露出分类过粗、准确性降低的情况。传统文献分类研究多以一级学科或二级学科等大类别分类或主题词分类为主，但是对学者而言，他们更需要细粒度分类。这种细粒度分类通常表现在两方面：一是对学科方向（领域）进一步划分，如在电子商务学科中进一步细分为跨境电商、农村电商、电商技术等；二是分类维度不仅有“研究内容”，还要有“研究方法”，如某篇文献按照研究对象被分为“跨境电商”同时按照研究方法分为“实证研究”，即对每篇学术文献进行双标签分类。

另一方面，现有学术文献分类的数据项主要来自题名、摘要、关键词等显性题录信息，然而，期刊名称、作者、研究机构等数据项与研究内容、研究方法之间虽然不存在显性相关性，但同一期刊、作者或者研究机构通常更聚焦特定研究内容、研究方法，可能存在隐性关系。因此，探索这种隐性关系，并以此提升学术文献分类的精度，成为本研究的另一目标。

鉴于此，本文以深度学习算法为基础，旨在探索一种基于题录信息的领域文献细粒度分类方法。题录信息作为分类标签的特征项，不仅包括与分类标签直接相关的题名、摘要、关键词，还包括与分类标签没有直接关系的期刊名称、作者以及研究机构等隐形特征。

2　文献综述

现有学术文献分类研究中，大部分学者是基于题录信息中的摘要、关键词和题名来进行文献分类的：武建光等把摘要中的高频特征词与人工识别的重点词作为中心词生成知识元，并将文献表征为若干个知识元，通过计算知识元的相似度进行文献分类[3]；H. Chu 和 Q. Ke 运用扎根理论方法对摘要中收集到的技术名称进行编码，以达到对研究方法分类的目的[4]；V. Chakraborty 等针对会计学术文献，将关键词和摘要作为原始数据，构建“文献－术语”矩阵以表示术语在文献中出现的频率[2]；周丽红和刘勘基于词性提取题名

* 本文系国家自然科学基金项目“作者、期刊与数据库错误引文的科学计量学研究：识别方法、产生机理与抑控对策”（项目编号：71603073）和河南省高校哲学社会科学创新团队资助项目“大数据与管理决策”（项目编号：2019－CXTD－04）研究成果之一。

作者简介：雷兵（ORCID：0000－0002－1073－4724），教授，博士；刘小（ORCID：0000－0002－6770－7583），硕士研究生；钟镇（ORCID：0000－0001－6248－2226），副教授，博士，通讯作者，E-mail：zhongzhen@haut.edu.cn。

收稿日期：2020－11－05　修回日期：2021－02－25　本文起止页码：128－137　本文责任编辑：王传清

雷兵，刘小，钟镇. 基于题录信息的领域学术文献细粒度分类方法研究[J]. 图书情报工作,2021,65(14):128－137.

和摘要中的特征词并进行筛选,然后计算特征词词频,将文献表征为特征向量的形式并通过关联规则进行文献分类[5];李慧和玄洪升根据专利文献的标题和摘要数据构建"文档－主题"矩阵和"主题－特征词"矩阵来挖掘技术创新主题[6]。除此之外,也有一部分学者尝试借助外部资源构建特征,以提升分类效果:李湘东等利用知网、维基百科、新闻页面等外部特征信息以提高文献分类的精度[7-9];苏燕等以医学主题词表(Medical Subject Headings,MeSH)为基础,筛选干细胞领域主题词作为表征文献的特征向量[10];潘东华等选用专利分类码构建技术词典,将文献表征为基于德温特手工代码(Derwent Manual Code,DMC)形式的向量,构建"专利－手工代码"矩阵,形成专利文献的技术知识图谱[11]。

使用机器学习算法提升文献分类精度是目前采用的主要技术路线,常用的机器学习算法有支持向量机(Support Vector Machine,SVM)、朴素贝叶斯(Naive Bayesian Model,NBM)、K－邻近模型(K-Nearest Neighbor,KNN)等,这些算法在文本分类领域中取得了较好的效果:S. Baker 等采用基于支持向量机的算法,对大规模的医学文献进行语义分类,得到了较高的准确度[12];L. Jiang 等采用局部加权的方式对朴素贝叶斯算法进行改进,从而提高了分类性能[13];白小明和邱桃荣对比了 KNN 与 SVM 算法对于科技文献分类的性能[14]。近年来,随着深度学习算法的日益成熟,已有研究表明卷积神经网络(Convolutional Neural Networks, CNN)可以自动从文本中学习特征,减少人工对特征信息的干预,使得文本分类的效果通常优于传统机器学习算法[15]:B. J. Gutierrez 等[16]使用多种机器学习算法对领域文献进行分类,验证了深度学习算法性能优于传统机器学习算法;郭利敏[17]基于 CNN 模型对《全国报刊索引》的 170 万条文献信息进行了多层次分类,取得了较好的效果。

综上所述,学者们基于摘要、关键词和题名等题录信息,采用机器学习算法对学术文献分类进行了深入研究,准确率不断提高。题录信息除了摘要、关键词和题名之外,还包括刊名、作者、机构等,然而,目前对此相关内容的研究鲜有涉及。本文在预研中也发现,直接将刊名、作者、机构等数据项加入特征向量中,无论采用传统机器学习还是深度学习算法,其分类准确率不但未提高,反而显著下降。分析其原因,主要是从摘要、关键词和题名中可以提取到与研究内容和研究方法相关的主题词,而刊名、作者、机构中几乎没有能直接表征文献特征信息的主题词,不加处理的在特征向量中引入这些数据项反而会导致更多的"噪声",降低分类精度。而事实上,每种期刊都对研究领域进行界定,且"偏爱"某些研究方法,每位作者有自己的研究领域和擅长的研究方法,每个研究机构或研究团队也会形成特定研究领域和常用的研究方法。本文推断这些数据项中存在着与研究内容和研究方法有关的隐性特征。

因此,本研究将题录信息数据项划分为显性特征和隐性特征,其中,摘要、关键词和题名为显性特征,刊名、作者及机构为隐性特征。对于显性特征,直接提取特征词;对于隐性特征,进行特征映射,将其显性化。在此基础上构建特征词数组,并使用 Skip-Gram 构建词向量模型,作为深度学习 CNN 模型的输入数据。在 CNN 模型的输出层,本文实现了同时对"研究内容"和"研究方法"的双标签分类。最后,本文通过电子商务主题领域文献验证方法的有效性。

3 模型构建

本研究预先对训练集和测试集中的学术文献,按照"研究内容"和"研究方法"两种分类分别进行人工标注,作为后续机器学习的语料。

领域学术文献主题分类的基本思路是:以题录信息作为分类依据,在特征提取的基础上构建初始特征矩阵,并对其进行词向量化,随后通过 CNN 深度学习算法实现细粒度分类,即按"研究内容"和"研究方法"的双标签分类。学术文献的题录信息一般包括题名、作者、机构、刊名、关键词以及摘要等,如表 1 所示:

表 1 学术文献题录信息示例

数据项	内容示例
题名	论电商平台"二选一"行为的法律规制
作者	XXX
机构	XX 大学法学院
刊名	现代法学
关键词	数字经济;电子商务;平台"二选一";P2B 条例
摘要	与传统经济一样,数字经济背景下的强制性"二选一"行为不是"本身违法",但如果行为人使用这种手段,严重损害竞争对手实现最低规模经济的能力,或者阻止新企业进入市场,就会在很大程度上妨碍市场竞争。考虑到进入市场存在着经济、技术、数据等各种障碍,特别是网络外部效应,我国电商平台已经高度集中。为了维护市场的竞争性,使商户和消费者充分感受电子商务的好处和便利,竞争执法机关应当保证平台商户的多归属,即任何平台经营者都无权强迫商户只能在一个平台上交易。同时,考虑到电子商务的特点和中小商户对平台中介的依赖性,我国有必要制定规范中介平台与商户之间交易关系的专门法,并完善《电子商务法》第 35 条。

本文提出的分类模型主要包括 3 个部分:①特征词典与停用词典的构建。选择训练集中所有文献的题

名、摘要和关键词构建"本地特征词词典"(user_dict)与"停用词词表"(stop_words list),以提高分词的准确性。②特征矩阵构建及向量化。将题录信息划分为显性特征(摘要、题名、关键词)和隐性特征(作者、刊名和机构)。对于显性特征进行分词、去停用词处理;对于隐性特征,采用特征映射的方式,将隐性特征显性化,这也是本研究的核心所在。在此基础之上构建基于题录信息的特征词数组并进行向量化处理,作为 CNN 分类模型的输入数据。③文献分类的深度学习。通过 CNN 模型对文献进行分类,在 CNN 模型的输出层,设计了"C(研究内容)×M(研究方法)"的形式,用以实现文献的双标签分类。如图 1 所示:

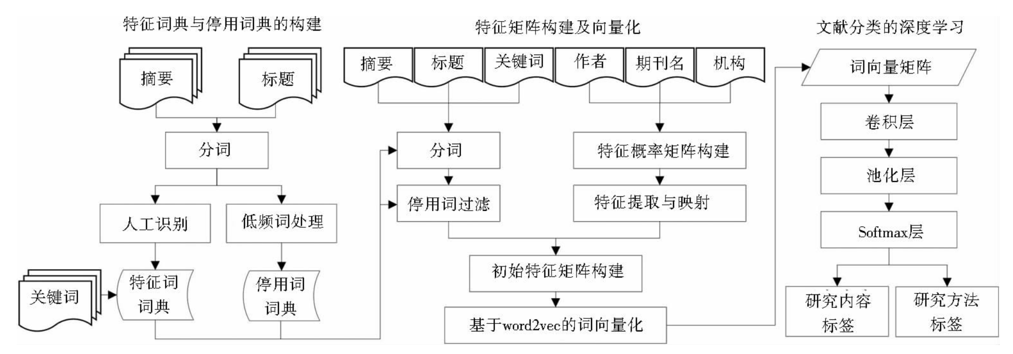


图 1　领域学术文献主题分类方法框架

3.1　停用词与特征词词典的构建

特征词词典与停用词词典的构建是数据预处理的重要环节,特征词词典作为分词的自定义词典参与到自动分词中,可以提高分词的准确性。停用词词典有助于过滤分词结果中的"噪声",从而提高深度学习模型的效果,避免出现过拟合现象。针对领域文献,本研究设计一种基于训练集所有文献题名、摘要和关键词数据项的特征词与停用词词典构建方法,具体说明如下:

特征词词典主要是由 3 个部分组成:首先,考虑到题录信息中关键词的重要性[18],将其全部纳入特征词词典。其次,题名和摘要中的高频词(大于等于 5)纳入特征词词典。最后结合领域专家的知识,将表征领域文献主题的典型词汇纳入特征词词典。

对于停用词词典,本研究最初仅采用哈工大停用词词库[19],但效果不佳。经分析发现,导致这种情况的原因主要有两点:一是学术文献的形式化描述词汇较多,如"随着""指出""按照"等句首词,这些词汇容易"误导"机器学习。二是分类特征不显著的低频词(小于 5),在机器学习中容易出现过拟合现象。基于此,本研究的停用词词典除包含哈工大停用词词库外,还加入了句首词和分类特征不显著的低频词。

3.2　特征矩阵的构建及向量化

在文本挖掘领域,深度学习模型可以自动从分布式词向量中寻找特征,相对于传统机器学习算法如条件随机场、支持向量机等,可移植性强、学习效率高[20]。但是初始特征的质量仍然影响深度学习效率,质量较差的特征容易出现过拟合或欠拟合现象。本研究针对题录信息中的数据项,将其划分为显性特征和隐性特征,分别进行处理。

3.2.1　显性特征提取

首先,将关键词直接加入到关键词特征集合 K 中;其次,引入特征词与停用词词典,使用分词工具对题名和摘要进行分词,形成题名特征集合 T、摘要特征集合 S。如式(1)-式(3)所示:

$$K=(k_1,k_2,\cdots,k_r) \quad 式(1)$$

$$T=(t_1,t_2,\cdots,t_p) \quad 式(2)$$

$$S=(s_1,s_2,\cdots,s_q) \quad 式(3)$$

其中,k_r 表示关键词中的第 r 个词,t_p 表示题名中的第 p 个词,s_q 表示摘要中的第 q 个词。这里需要特别说明的是:每篇文献对应的 r、p、q 是可变的,为了使后续词向量的长度固定,需要设置 3 个超参数 $R(\geq r)$、$P(\geq p)$、$Q(\geq q)$,用于固定 K、T、S 的长度,不足部分用"0"占位,具体形式见式(4)-式(6):

$$K=(k_1,k_2,\cdots,k_r,\overbrace{0,\cdots,0}^{R-r}) \quad 式(4)$$

$$T=(t_1,t_2,\cdots,t_p,\overbrace{0,\cdots,0}^{P-p}) \quad 式(5)$$

$$S=(s_1,s_2,\cdots,s_q,\overbrace{0,\cdots,0}^{Q-q}) \quad 式(6)$$

3.2.2　隐性特征映射

文题录信息中的机构、刊名、作者与文献的研究内容或研究方法存在隐性关联。以电商领域为例:计算机学院发表的学术论文可能聚焦于"信息技术的电子商务应用",而法学院发表的论文围绕"电子商务法律

雷兵，刘小，钟镇. 基于题录信息的领域学术文献细粒度分类方法研究[J]. 图书情报工作,2021,65(14):128－137.

法规”展开；农业类期刊则可能刊发“农村电商扶贫”主题的学术论文，国际贸易类期刊可能选择“跨境电商”的主题进行讨论同理，特定的作者通常使用相对固定的研究方法且聚焦某一研究领域，但当该学者与其他学者合作时，其研究内容或研究方法又可能改变。因此，本研究采用特征映射的方式将隐含在机构、刊名、作者中的特征信息显性化，然后加入到初始特征矩阵中，具体映射过程如下：

（1）作者特征处理。将作者与领域文献进行关联，根据作者与发表文献所涉及研究内容、研究方法的共现频次，将作者这个隐性特征进行显性化处理。图2展示了不同类型作者的研究方法标签生成过程：

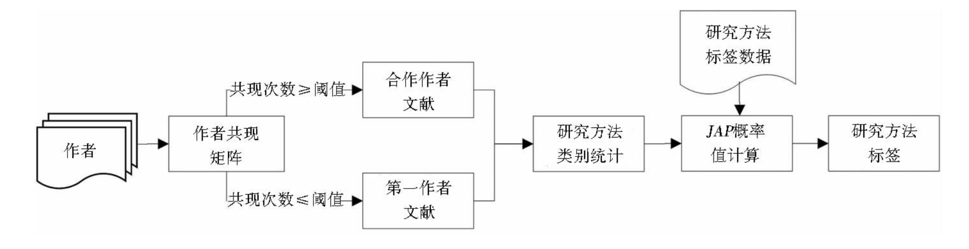


图2　作者映射流程

根据领域文献的作者合著关系构建作者共现矩阵，如果共现次数（即合著频次）超过特定阈值，则认为两个作者在某一领域的学术研究中存在稳定的合作关系，将其视为合作作者进行特征映射，否则对第一作者进行特征映射。具体做法是：首先对合作作者（或第一作者）按照研究方法类别进行频数统计，生成“作者－研究方法”频数分布表。然后，计算不同作者采用的研究方法概率值 *JAP*，计算公式见式（7）。*JAP* 值越大，表明某一作者对某一研究方法的偏好越强，最后，生成“作者－研究方法”概率分布表。

$$JAP_i(j) = \frac{m_{ij}}{\sum_{i=1}^{M} m_{ij}} \quad \text{式(7)}$$

式（7）中，*M* 表示领域文献研究方法的类别，m_{ij} 表示作者 j 采用第 i 个研究方法的频数。根据概率分布表，将作者映射为研究方法显性特征：首先设置 *JAP* 转化概率阈值，然后选择 *JAP* 值最大并且满足阈值的研究内容标签，并将作者映射为此标签。阈值为超参数，假设经过试验，将阈值设置为 0.7，即如果某作者的 *JAP* 值不低于 0.7，就将该作者映射为该研究方法标签，否则用占位符“0”代替。表2显示了作者映射为研究方法标签的示例，“作者－1”映射为“研究方法－2”，“合作作者－2”映射为“研究方法－1”，“作者－5”映射为“研究方法－3”，而“作者－3”“合作作者－4”用占位符“0”代替。

表2　“作者－研究方法”概率分布表示例

	研究方法－1	研究方法－2	研究方法－3	研究方法－4
作者－1	0	1	0	0
合作作者－2	0.8	0.2	0	0
作者－3	0	0	0.5	0.5
合作作者－4	0.6	0	0.4	0
作者－5	0	0	1	0

（2）期刊特征处理。与作者特征处理类似，将刊名与领域文献关联起来，映射为研究内容、研究方法显性特征。以研究内容为例，处理流程如图3所示：

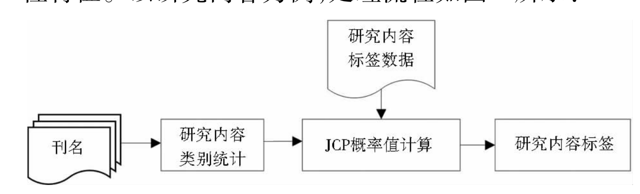


图3　期刊映射流程

首先，以刊名为对象，统计每种期刊不同研究内容的频数，生成“期刊－研究内容”频数分布表，并计算每种期刊“研究内容”标签的概率值 *JCP*，计算公式如式（8）所示。同样，*JCP* 的值越大，表示某一种期刊对于某一研究内容的偏好越强。然后根据 *JCP* 生成“期刊－研究内容”概率分布表。

$$JCP_i(j) = \frac{c_{ij}}{\sum_{i=1}^{C} c_i j} \quad \text{式(8)}$$

式（8）中，*c* 表示领域文献的研究内容类别，c_{ij} 表示期刊 j 对于第 i 个研究内容标签的频数。根据概率分布表，将刊名映射为研究内容显性特征：首先设置 *JCP* 转化概率阈值，然后将刊名转化为大于等于阈值的研究内容标签。如果没有满足条件的标签或标签数量不足，则用占位符“0”代替。假设经过试验，将阈值设置为 0.33，如果某一研究内容 *JCP* 值不低于 0.33，则将该研究内容标签加入到刊名映射集合中。表3显示期刊名称映射为研究内容标签的实例。“期刊－1”的映射集合为{研究内容－1，研究内容－4，0}，“期刊－3”的映射集合为{研究内容－3，研究内容－4，研究内容－5}，“期刊－5”的映射集合为{0，0，0}。

表 3 “期刊 – 研究内容”概率分布表示例

	研究内容 –1	研究内容 –2	研究内容 –3	研究内容 –4	研究内容 –5
期刊 –1	0.60	0	0	0.40	0
期刊 –2	0.15	0.10	0.75	0	0
期刊 –3	0	0	0.34	0.33	0.33
期刊 –4	0	1	0	0	0
期刊 –5	0.25	0	0.20	0.25	0.30

（3）研究机构特征处理。将研究机构与领域文献关联起来，映射为研究内容、研究方法显性特征。以研究内容为例，其处理流程见图 4。

首先对研究机构进行如下处理：①若文献中存在多个研究机构，仅选取第一研究机构；②通过正则表达式来进行一级研究机构与二级研究机构的划分，考虑到一级研究机构如“XX 大学”基本无法表明领域文献的研究内容，因此，对一级研究机构进行删除仅保留二级研究机构，如经济与管理学院、法学院进行特征映射。之后，计算每个研究机构中出现的“研究内容”与“研究方法”标签的概率值，并采取与期刊映射相同的方法进行特征映射。

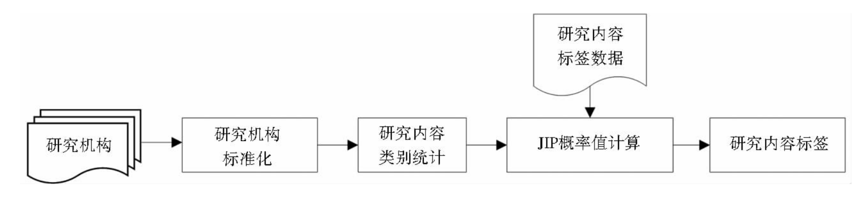


图 4 研究机构映射流程

3.2.3 词向量化

将处理后的显性特征和隐性特征分别加入到特征词数组 D 中，如式（9）所示。

$$D=[K,T,S,A,J,O] \quad \text{式(9)}$$

其中，K、T、S、A、J、O 分别表示经过处理后的题名、关键词、摘要、作者、刊名、机构数据。

然后，通过 Word2Vec 将 D 转化为词向量化，形成后续深度学习模型的初始化特征矩阵。Word2Vec 是一种浅层的神经网络模型，是单词在多维数字空间的映射，数字空间的位置表明了单词的语义信息[21]。Skip-Gram 是 Word2Vec 词向量模型的一种方法，可以通过中心词预测上下文词出现的概率，使用预训练的词向量会使 CNN 模型的分类性能得到明显提升[22]。借鉴 A. Timoshenko 等[23]对 Skip-gram 模型的参数进行设置，本文将滑动窗口 c 的大小设置为 5，词向量维度 d 设置为 20，将数组 D 输入到词向量模型中，输出为词向量矩阵 $D^* \in i^{d\times n}$，并以此作为 CNN 模型的输入。

3.3 文献分类的深度学习

与传统的机器学习算法相比，深度学习模型在大规模的文本分类领域取得了较好的性能[17,24]。深度学习模型通过神经元的连接，可以从浅层的初级特征开始学习到深层的高级特征。对于本研究所构建的词向量矩阵 D^*，深度学习模型 CNN 既可以学习到全局特征，又可以学习到不同题录信息所包含的细节特征。

CNN 模型是由输入层、卷积层、池化层与 Softmax 层组成，并利用梯度下降方法对权重参数反向调节[25]，具体结构见图 5。CNN 的输入层为词向量矩阵 D^*，卷积层通过若干卷积核对初始特征矩阵进行卷积操作，形成特征图。之后，对特征图进行池化操作，以减少维度并保留最大的特征值。池化层可以过滤掉无用特征，保留重要特征。Softmax 层通过全连接的方式将池化层输出的向量经过 Softmax 函数转换为文献主题的概率值，用以预测文献类别。

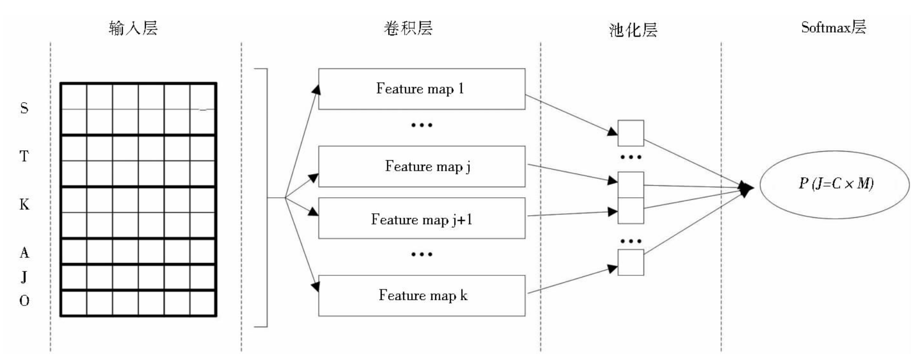


图 5 CNN 模型结构

雷兵，刘小，钟镇. 基于题录信息的领域学术文献细粒度分类方法研究[J]. 图书情报工作,2021,65(14):128－137.

对于研究主题集合 J,本文将“研究内容”与“研究方法”同时作为 CNN 模型的输出,组合方式如式(10)所示:

$$J = C \times M \quad 式(10)$$

其中,C 表示研究内容标签集合,M 表示研究方法标签集合,J 表示研究内容标签与研究方法标签的组合,从而实现双标签分类,具体过程见图 6。如假设某领域文献有 4 种研究方法、8 类研究内容,主题标签就需要设置 32 个,分别为主题标签 1、主题标签 2,直到主题标签 32。若某篇文献标注为“主题标签 32”,说明它的研究内容、研究方法分别是“研究内容 8”“研究方法 4”。

图 6　文献双标签分类实现过程

4　实验验证

为验证上述分类模型的可行性与有效性,本研究对中国知网(CNKI)数据库中的“电子商务”主题文献进行研究内容与研究方法分类,并与支持向量机、朴素贝叶斯等传统机器学习算法进行对比。

4.1　数据来源

本研究的数据来自 CNKI 的中国期刊全文数据库,以“电子商务”为检索词进行主题检索,期刊类别限定为 CSCD、EI 与 CSSCI,检索时间范围为 1998 年 5 月 15 日－2020 年 6 月 10 日,共计检索到 8 874 条记录,下载的内容包括题名、研究机构、出版期刊、关键词、摘要等信息。经过去重、噪声处理与缺失值处理,最终确定了 7 647 篇待标注文献。

在 7 647 篇文献中,共包含 13 977 个关键词,出现频次最高的特征词为“跨境电子商务”,共计 136 次;共包含 785 种期刊,其中单一期刊出现的最高频次为 291 次(中国流通经济),560 种期刊出现的频次在 2 次以上;共包含 6 785个研究机构,单一研究机构最高发表频次为 82 次(武汉大学信息管理学院),1 899 个研究机构出现的频次在 2 次以上;共包含 10 568 位作者,其中单一作者出现的最高频次为 33 次,2 262 位作者出现的频次在 2 次以上。高频主题词、期刊与研究机构的分布情况如表 4 所示:

表 4　高频主题词、期刊、研究机构频数分布情况

主题词	频次	期刊	频次	研究机构	频次
跨境电子商务	136	中国流通经济	291	武汉大学信息管理学院	82
企业管理	131	情报杂志	290	吉林大学管理学院	77
信息化	109	情报科学	233	华中科技大学管理学院	75
推荐系统	77	图书情报工作	179	西安交通大学管理学院	68
供应链	73	科技管理研究	166	重庆大学经济与工商管理学院	66
信息技术	72	计算机工程	133	西安交通大学经济与金融学院	65
网络营销	68	现代情报	115	上海理工大学管理学院	57
协同过滤	68	商业研究	115	北京邮电大学经济管理学院	52
商业模式	67	生产力研究	106	中国人民大学商学院	46
物流	63	科技进步与对策	85	复旦大学管理学院	44

4.2　人工标引

通过前期文献调研,借鉴肖连杰和章成志等[20,26]的领域文献分类方法,并与多名电商学者反复讨论,最终确定 13 个分类标签,分为“研究内容”与“研究方法”两个大类,覆盖了目前电子商务研究的主要研究领域,如表 5 所示:

表 5　领域文献主题标签

分类	主题标签	数量
研究内容	商业模式;法律法规;物流配送、支付、金融;市场营销;电商技术;农村电商;跨境电商;信用风险;电商其他	9
研究方法	理论研究;实证研究;案例研究;技术研究	4
总计	13	

在表 5 的研究内容主题标签中,"电商其他"表示电子商务比较小众的研究领域,如电商人才培养等。在研究方法上,本研究将主要采用定性研究的方式分析概念或解读政策的研究归类为"理论研究";将主要采用计量经济学或产业经济学方法研究或检验宏观与中观层面截面或时间序列数据的研究归类为"实证研究";通过构建模型并使用具体数据解析具体案例,一般聚焦于组织行为的研究归类为"案例研究";"技术研究"则指的是运用计算机技术对电子商务相关领域进行研究。

考虑到领域文献的专业性与标引标准的统一性[27],本文在人工标引部分采用人数较少的领域专家标引方法,而非业界流行的众包模式[28]。众包模式下的数据标引工作往往由众多非领域人员完成,虽然提高了标引的效率,但是不适合专业性较高的学术文献标引工作。在具体操作上,本文主要是基于领域专家的帮助,为每一个研究内容标签确定若干特征词(见表 6),并结合特征词的出现位置和频数进行研究内容标注。此外,若同一篇文献出现两种以上的研究内容,则根据特征词的频率进行标注,选取特征词频率高的研究内容标签。

表 6 电子商务领域文献研究对象标引特征

研究内容标签	主要特征词
商业模式	B2B 模式、B2C 模式、线上与线下融合等
法律法规	税法、电子商务法、消费者权益保护法等
物流配送、支付、金融	支付系统、物流配送、P2P 等
市场营销	在线评论、定价研究、消费者偏好等
电商技术	推荐算法、云计算等
农村电商	农产品电商、电商扶贫等
跨境电商	WTO、贸易便利化等等
信用分险	信任危机、可信度等
电商其他	其他电商领域研究问题

表 7 列出了 5 篇文献标引的例子。以第 3 篇为例,题录、关键词和摘要中均出现了与"农村电商"和"物流配送"相关的特征词,但是"物流配送"相关特征词出现了 5 次,"农村电商"相关特征词出现了 3 次,因而归为"物流配送、支付、金融"。而在研究方法上,题录和关键词中出现了"改进算法"字样,因此研究方法标签确定为"方法研究"。

表 7 标引示例

编号	题录	关键词	摘要	研究对象标签	研究方法标签
1	罗 XX,XX 大学信息学院,一种新型的匿名公平电子商务协议,《计算机应用研究》,2010	电子商务,信息安全,素质测评,指标	针对电子商务中客户、商家和第三方的信任问题设计了一种简单、有效的匿名公平电子商务协议,……	电商技术	方法研究
2	张 XX,XX 大学商学院,共生抑或迭代:再议跨境电子商务与全球数字贸易,《当代经济管理》,2020	跨境电子商务,全球数字贸易,数字技术	通过梳理数字贸易的已有概念界定,结合新时代发展背景,对全球数字贸易……	跨境电商	理论研究
3	盛 XX,XX 大学经济与管理学院,基于共同配送策略的农村电商集送货一体化车辆路径问题,《系统工程》,2019	共同配送,农村电商,集送货一体化车辆路径问题,改进蚁群算法	研究农村电商物流配送问题,综合考虑区域内多配送中心、客户居住地较分散、同时具有集货和送货双重需求……	物流配送、支付、金融	方法研究
4	EC-CDIO 电子商务张 XX,XX 大学管理科学与工程学院,人才培养模式的构建,《高等工程教育研究》,2019	电子商务,EC-CDIO,人才培养模式	基于 CDIO 人才培养模式,结合电子商务实践对人才素养的综合要求,提出并实践了……	电商其他	理论研究
5	王 XX,XX 大学,上海参与"两带一路"建设的优势、挑战与对乡村振兴战略背景下农村电商创业的典型模式研究——以江苏省创业实践为例,《农业经济与管理》,2019	乡村振兴,农村电商创业模式,农村电商创业要素	乡村振兴战略是新时代中国特色"三农"发展战略体系的重要组成部分,战略的实施为农村电商创业发展注入新动能,推动农村电商创业模式不断创新……	农村电商	案例研究

为确保标引结果的准确性,标引工作分别由 3 位电子商务研究方向的硕士研究生按照上述标引规则独立完成。如果两位及以上工作人员的标引结果一致,则确定该条文献的标签类别;如果 3 位工作人员的标引结果都不一致,则将该条文献交由领域专家处理。

4.3 实验分析

4.3.1 评价标准

领域文献主题分类使用准确率 P 值、召回率 R 值以及 F_1 值进行评估。本文采用的方法是将需要评估的类标签单独视为正类,其他类别视为负类,构建混淆矩阵对每个类别标签进行计算。被正确划分某一类别标签的样本数为 TP,被错误划分为这一标签的样本数为 FP,被正确划分到其他类别标签的样本数为 TN,被错误划分到其他类别标签的样本数为 FN,则 P、R 和 F_1 值分别为:

$$P = \frac{TP}{TP + FP} \quad 式(10)$$

$$R = \frac{TP}{TP + FN} \qquad \text{式(11)}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \qquad \text{式(12)}$$

4.3.2 对比分析

电子商务领域文献分类结果见表8、表9、表10,表8给出了本方法分类精度,表9为基于不同初始特征构建的分类精度,表10显示了基于其他机器学习模型的分类精度。

通过表8可以发现:在电子商务领域文献的分类结果中,对于研究内容,“农村电商”识别的准确率最高,达到了97%,对“商业模式”的分类准确率最低,为48%,“电商其他”“物流配送、支付、金融”的准确率也相对较低,其他标签的准确率都在70%以上。通过分析发现,分类结果不佳的研究内容,文献研究的范围相对较广,如标签为“电商其他”的文献,研究内容会涉及“电子商务人才培养”“旅游电商”等一些覆盖率较低的研究内容。研究内容的不一致性导致了“商业模式”与“电商其他”类中的文献特征离散化程度较高,分类结果相对较差。对于研究方法,除“案例研究”分类准确率较低外,其他研究方法的准确率都在85%以上,分类结果较好。对于案例研究,通过统计分析发现,电商领域中使用案例研究方法的文献比例较小,在所有标注的文献中仅占7.36%,较少数量的文献可能导致模型对“案例研究”方法的特征提取效果较差,出现过拟合现象。

表8　电子商务领域文献主题类别标签分类结果

研究主题		性能指标		
类别	标签	P	R	F_1
研究内容	商业模式	0.54	0.64	0.59
	法律法规	0.84	0.84	0.84
	物流配送、支付、金融	0.58	0.74	0.65
	市场营销	0.74	0.64	0.69
	电商技术	0.76	0.80	0.78
	农村电商	0.98	0.88	0.90
	跨境电商	0.97	0.92	0.94
	信用风险	0.79	0.60	0.68
	电商其他	0.61	0.59	0.60
研究方法	理论研究	0.92	0.96	0.94
	实证研究	0.89	0.85	0.87
	方法研究	0.85	0.72	0.78
	案例研究	0.62	0.69	0.65

为检验本研究所构建的初始化特征矩阵有用性,本文采用不同的方法进行对比实验,实验以本研究所提出的研究方法为基础,每次只改变一个特征项,其他

表9　不同数据输入与预处理的电子商务领域文献主题分类对比结果

输入输数与预处理	性能指标					
	研究内容			研究方法		
	P	R	F_1	P	R	F_1
本文分类模型	0.72	0.73	0.74	0.88	0.80	0.81
将刊名、作者、机构名直接加入	0.63	0.62	0.62	0.75	0.70	0.72
仅题名、摘要数据	0.71	0.72	0.72	0.79	0.78	0.78
仅“哈工大”停用词词典	0.69	0.70	0.70	0.77	0.78	0.77

表10　不同模型的电子商务领域文献主题分类结果

模型	性能指标					
	研究内容			研究方法		
	P	R	F_1	P	R	F_1
本文分类模型	0.72	0.73	0.74	0.88	0.80	0.81
SVM	0.57	0.60	0.58	0.69	0.41	0.51
NBM	0.64	0.67	0.65	0.70	0.67	0.68
KNN	0.50	0.50	0.50	0.69	0.45	0.54

特征项不变。结果见表9:基于题录信息的文献分类模型对研究内容分类的准确率为72%,召回率为73%,宏 F_1 值为74%,对研究方法分类的准确率为88%,召回率为80%,宏 F_1 值为81%。将作者、机构、刊名原始数据直接加入到特征矩阵中,研究内容与研究方法的宏值 F_1 分别下降了9%和11%。然后采用其他学者运用的CNN文献分类算法仅使用“题名”和“摘要”数据进行实验[17],研究内容与研究方法的宏 F_1 值与本文的研究方法2%和3%。在数据预处理过程中,如果不加入领域特征词典,并仅以“哈工大停用词”来对原始数据分词,研究内容与研究方法分类结果的宏 F_1 值都相差了4%。这直接表明,本文对作者、机构和刊名进行特征映射、构建初始特征矩阵对于提升模型分类效果有一定的帮助。

为验证CNN算法对领域文献细粒度分类的有效性,在进行电子商务领域文献分类时,本文将常见的机器学习算法作为对比实验,包括经典的支持向量机算法、朴素贝叶斯算法以及K－邻近模型算法[29－30]。在实验中,除模型不同外,其他特征项均相同,实验结果见表10。由表10可见,使用本文分类模型(基于CNN算法)的分类效果最佳。在传统机器学习算法中,NBM算法用于文献主题分类的效果表现较好,但是相对于CNN算法差距比较明显,研究内容分类结果的宏 F_1 值相差9%,研究方法相差13%。

通过表9、表10的对比分析可以看出,本研究所提出的方法提高了文献分类结果的宏 F_1 值,这说明对于领域文献主题细粒度分类问题,本研究所提出的方法

是有效的。

5 结语

本研究构建了基于题录信息的细粒度文献分类模型。首先对表征文献主题的题录信息进行筛选,并基于训练集所有文献题名、关键词和摘要构建特征词与停用词词典;然后对题录信息中的关键词、标题、摘要等显性特征进行特征提取,对作者、刊名、机构等隐形特征进行特征映射,构建特征数组;接着对特征数组进行词向量训练,并以此作为 CNN 模型的输入;最后通过 CNN 模型实现领域文献"研究内容"与"研究方法"的双标签分类。实验结果发现,本研究提出的分类模型不仅明显优于基于传统机器学习的分类方法,也比仅使用题目信息中显性特征的深度学习分类方法精度高。

基于题录信息的领域文献主题分类同样面临一些问题:第一是领域文献研究主题的复杂性,同一文献的研究主题可能包括多个方面,本文仅依据出现频数最高的特征词进行主题标签的划分,这就容易导致分类出现不准确的情况。如对于某一篇研究农村电商物流的文献,根据特征词的频数我们标引的主题为"物流配送、支付、金融",但是模型的分类结果为农村电商,这表明了单类别输出的缺陷。在后续研究中,将继续对 CNN 模型的输出层进行改进,设计多类别的输出以增强文献主题分类的准确性。第二是文献的分类标签依赖领域专家,存在一定的主观局限性,后续拟采用机器学习算法与领域专家知识相结合的方式来获取领域文献的研究主题,以提高模型自动化分类的能力。第三是本文实验所使用的数据规模比较小,今后将开展大规模领域文献分类实验,进一步验证本文所提出方法的有效性。

参考文献:

[1] 刘爱军, 俞立平. 文献计量指标的客观分类及其启示——以 JCR 2015 经济学期刊为例[J]. 情报理论与实践, 2017, 40(7): 33-37, 49.

[2] CHAKRABORTY V, CHIU V, VASARHELYI M. Automatic classification of accounting literature[J]. International journal of accounting information systems, 2014, 15(2): 122-148.

[3] 武建光, 苏云梅, 于琦, 等. 基于知识元的学术文献分类研究[J]. 情报理论与实践, 2019, 42(3): 160-165.

[4] CHU H, KE Q. Research methods: what's in the name? [J]. Library & information science research, 2017, 39(4): 284-294.

[5] 周丽红, 刘勘. 基于关联规则的科技文献分类研究[J]. 图书情报工作, 2012, 56(4): 12-16, 119.

[6] 李慧, 玄洪升. 专利视角下融合多属性的技术创新主题挖掘方法——以芯片领域专利为例[J]. 图书情报工作, 2020, 64(11): 96-107.

[7] 李湘东, 刘康, 丁丛, 等. 基于《知网》的多种类型文献混合自动分类研究[J]. 现代图书情报技术, 2016(2): 59-66.

[8] 李湘东, 阮涛, 刘康. 基于维基百科的多种类型文献自动分类研究[J]. 数据分析与知识发现, 2017, 1(10): 43-52.

[9] 李湘东, 高凡, 李悠海. 共通语义空间下的跨文献类型文本自动分类研究[J]. 数据分析与知识发现, 2018, 2(9): 66-73.

[10] 苏燕, 徐萍, 孔亮亮, 等. 基于 MeSH 的生物医学分类主题词表重构探索——以干细胞研究文献为例[J]. 图书馆杂志, 2015, 34(3): 47-52.

[11] 潘东华, 徐珂珂. 基于专利文献分类码的技术知识图谱绘制方法研究[J]. 情报学报, 2015, 34(8): 866-874.

[12] BAKER S, SILINS I, GUO Y, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer [J]. Bioinformatics, 2016, 32(3): 432-440.

[13] JIANG L, CAI Z, ZHANG H, et al. Naive Bayes text classifiers: a locally weighted learning approach[J]. Journal of experimental & theoretical artificial intelligence, 2013, 25(2): 273-286.

[14] 白小明, 邱桃荣. 基于 SVM 和 KNN 算法的科技文献自动分类研究[J]. 微计算机信息, 2006(36): 275-276, 265.

[15] WANG S, HUANG M, DENG Z. Densely connected CNN with multi-scale feature attention for text classification[C]//Proceedings of the 27th international joint conference on artificial intelligence (IJCAI). Stockholm: International Joint Conferences on Artificial Intelligence Organization, 2018: 4468-4474.

[16] GUTIERREZ B J, ZENG J, ZHANG D, et al. Document classification for COVID-19 literature[BE/OL]. [2020-09-04]. https://arxiv.org/abs/2006.13816.

[17] 郭利敏. 基于卷积神经网络的文献自动分类研究[J]. 图书与情报, 2017 (6): 96-103.

[18] 杜德慧, 李长玲, 相富钟, 等. 基于引文关键词的跨学科相关知识发现方法探讨[J]. 情报杂志, 2020, 39(9): 189-194.

[19] 俞琰, 赵乃瑄. 基于辅助集的专利主题分析领域停用词选取[J]. 数据分析与知识发现, 2018, 2(11): 95-103.

[20] 肖连杰, 孟涛, 王伟, 等. 基于深度学习的情报分析方法识别研究——以安全情报领域为例 [J]. 数据分析与知识发现, 2019, 3(10): 20-28.

[21] MARCO B, GEORGIANA D, GERMAN K. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors [C]//Proceedings of the 52nd annual meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014: 238-247.

[22] ZHANG Y, WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. Computer science, 2015(10): 253-263.

[23] TIMOSHENKO A, HAUSER J R. Identifying customer needs from user-generated content[J]. Marketing science, 2019, 38(1): 1

－20.
[24] YAN Y, YIN X-C, YANG C, et al. Biomedical literature classification with a CNNs-based hybrid learning network[J]. Plos one, 2018, 13(7): 1－31.
[25] KIM Y. Convolutional neural networks for sentence classification [BE/OL]. [2020－09－07]. https://arxiv.org/abs/1408.5882.
[26] 章成志, 李卓, 储荷婷. 基于全文内容的学术论文研究方法自动分类研究[J]. 情报学报, 2020, 39(8): 852－862.
[27] 唐琳, 郭崇慧, 陈静锋,等. 基于中文学术文献的领域本体概念层次关系抽取研究[J]. 情报学报, 2020, 39(4): 387－398.
[28] WILLIS C G, LAW E, WILLIAMS A C, et al. CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens[J]. New phytologist, 2017, 215(1): 479－488.
[29] 张华鑫,庞建刚. 基于 SVM 和 KNN 的文本分类研究[J]. 现代情报,2015,35(5):73－77.
[30] 萧莉明,于宽,蔡珣. 一种基于 Bayes 分类器的中文期刊自动分类系统[J]. 现代情报,2007(4):146－147,150.

作者贡献说明:
雷兵:研究设计与论文修改;
刘小:模型构建、数据检验与论文撰写;
钟镇:选题与论文修改。

Research on Fine-Grain Classification Method of Academic Literature Based on Bibliographies

Lei Bing[1,2] Liu Xiao[1,2] Zhong Zhen[1,2]
[1] School of Management, Henan University of Technology, Zhengzhou 450001
[2] Business Intelligence and Knowledge Engineering Laboratory, Henan University of Technology, Zhengzhou 450001

Abstract: [**Purpose/significance**] Targeting the academic literature in a specific field, a dual classification model in "research content" and "research method" is constructed based on bibliographies, aiming to provide method reference for fine-grain classification of academic literature. [**Method/process**] Using the convolutional neural network in deep learning as the basic model, the title, abstract, keyword, source, author, organ and other bibliographies were divided into dominant feature and invisible feature. Through dominant feature extraction, invisible feature mapping and other steps, a feature word array was formed. On this basis, the word vector matrix was constructed, which processed by the convolutional layer, pooling layer and Softmax layer to complete the classification task. [**Result/conclusion**] Take the literature in the e-commerce field as an example for experimental verification. The results show that the macro F_1 values of this model are 0.74 and 0.81 respectively according to the two categories of "research content" and "research method". The classification results are not only significantly better than traditional machine learning methods, but also higher than deep learning classification methods that only use dominant feature.

Keywords: academic literature subject classification bibliographies deep learning convolutional neural network